MOLECULAR DIFFERENCES BETWEEN SPECIES OF THE *M. TUBERCULOSIS* COMPLEX

[0001] Tuberculosis is an ancient human scourge that continues to be an important public health problem worldwide. It is an ongoing epidemic of staggering proportions. Approximately one in every three people in the world is infected with *Mycobacterium tuberculosis*, and has a 10% lifetime risk of progressing from infection to clinical disease. Although tuberculosis can be treated, an estimated 2.9 million people died from the disease last year.

[0002] There are significant problems with a reliance on drug treatment to control active *M. tuberculosis* infections. Most of the regions having high infection rates are less developed countries, which suffer from a lack of easily accessible health services, diagnostic facilities and suitable antibiotics against M. tuberculosis. Even where these are available, patient compliance is often poor because of the lengthy regimen required for complete treatment, and multidrug-resistant strains are increasingly common.

[0003] Prevention of infection would circumvent the problems of treatment, and so vaccination against tuberculosis is widely performed in endemic regions. Around 100 million people a year are vaccinated with live bacillus Calmette-Guerin (BCG) vaccine. BCG has the great advantage of being inexpensive and easily administered under less than optimal circumstances, with few adverse reactions. Unfortunately, the vaccine is widely variable in its efficacy, providing anywhere from 0 to 80% protection against infection with *M. tuberculosis.*

[0004] BCG has an interesting history. It is an attenuated strain of *M. bovis*, a very close relative of *M. tuberculosis*. The *M. bovis* strain that became BCG was isolated from a cow in the late 1800's by a bacteriologist named Nocard, hence it was called Nocard's bacillus. The attenuation of Nocard's bacillus took place from 1908 to 1921, over the course of 230 *in vitro* passages. Thereafter, it was widely grown throughout the world, resulting in additional hundreds and sometime thousands of *in vitro* passages. Throughout its many years in the laboratory, there has been

selection for cross-reaction with the tuberculin skin test, and for decreased side effects. The net results have been a substantially weakened pathogen, which may be ineffective in raising an adequate immune response.

[0005] New antituberculosis vaccines are urgently needed for the general population in endemic regions, for HIV-infected individuals, as well as health care professionals likely to be exposed to tubercle bacilli. Recombinant DNA vaccines bearing protective genes from virulent *M. tuberculosis* are being developed using shuttle plasmids to transfer genetic material from one mycobacterial species to another, for example see U.S. Patent no. 5,776,465. Tuberculosis vaccine development should be given a high priority in current medical research goals.

Relevant literature

[0006] Mahairas *et al.* (1996) J Bacteriol **178**(5):1274-1282 provides a molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis.* Subtractive genomic hybridization was used to identify genetic differences between virulent *M. bovis* and *M. tuberculosis* and avirulent BCG. U.S. Patent No. 5,700,683 is directed to these genetic differences.

[0007] Cole *et al.* (1998) Nature **393**:537-544 have described the complete genome of *M. tuberculosis*. To obtain the contiguous genome sequence, a combined approach was used that involved the systematic sequence analysis of selected large-insert clones as well as random small-insert clones from a whole-genome shotgun library. This culminated in a composite sequence of 4,411,529 base pairs, with a G + C content of 65.6%. 3,924 open reading frames were identified in the genome, accounting for ~91% of the potential coding capacity.

[0008] *Mycobacterium tuberculosis* (*M.tb.*) genomic sequence is available at several internet sites.

SUMMARY OF THE INVENTION

[0009] Genetic markers are provided that distinguish between strains of the *Mycobacterium tuberculosis* complex, particularly between avirulent and virulent

strains. Strains of interest include *M. bovis*, *M. bovis* BCG strains, *M. tuberculosis (M. tb.)* isolates, and bacteriophages that infect mycobacteria. The genetic markers are used for assays, *e.g.* immunoassays, that distinguish between strains, such as to differentiate between BCG immunization and *M. tb.* infection. The protein products may be produced and used as an immunogen, in drug screening, *etc.* The markers are useful in constructing genetically modified *M. tb* or *M. bovis* cells having improved vaccine characteristics.

DETAILED DESCRIPTION OF THE EMBODIMENTS

[0010] Specific genetic deletions are identified that serve as markers to distinguish between avirulent and virulent mycobacteria strains, including *M. bovis*, *M. bovis* BCG strains, *M. tuberculosis (M. tb.)* isolates, and bacteriophages that infect mycobacteria. These deletions are used as genetic markers to distinguish between the different mycobacteria. The deletions may be introduced into *M. tb.* or *M. bovis* by recombinant methods in order to render a pathogenic strain avirulent. Alternatively, the deleted genes are identified in the *M. tb.* genome sequence, and are then reintroduced by recombinant methods into BCG or other vaccine strains, in order to improve the efficacy of vaccination.

[0011] The deletions of the invention are identified by comparative DNA hybridizations from genomic sequence of mycobacterium to a DNA microarray comprising representative sequences of the *M. tb.* coding sequences. The deletions are then mapped to the known *M. tb.* genome sequence in order to specifically identify the deleted gene(s), and to characterize nucleotide sequence of the deleted region.

[0012] Nucleic acids comprising the provided deletions and junctions are used in a variety of applications. Hybridization probes may be obtained from the known *M. tb.* sequence which correspond to the deleted sequences. Such probes are useful in distinguishing between mycobacteria. For example, there is a 10% probability that an *M. tb.* infected person will progress to clinical disease, but that probability may vary depending of the particular infecting strain. Analysis for the presence or absence of

the deletions provided below as "*M. tb* variable" is used to distinguish between different *M. tb* strains. The deletions are also useful in identifying whether a patient that is positive for a tuberculin skin test has been infected with *M. tb* or with BCG.

[0013] In another embodiment of the invention, mycobacteria are genetically altered to delete sequences identified herein as absent in attenuated strains, but present in pathogenic strains, *e.g.* deletions found in BCG but present in *M. tb* H37Rv. Such genetically engineered strains may provide superior vaccines to the present BCG isolates in use. Alternatively, BCG strains may be "reconstructed" to more closely resemble wild-type *M. tb* by inserting certain of the deleted sequences back into the genome. Since the protein products of the deleted sequences are expressed in virulent mycobacterial species, the encoded proteins are useful as immunogens for vaccination.

[0014] The attenuation (loss of virulence) in BCG is attributed to the loss of genetic material at a number of places throughout the genome. The selection over time for fewer side-effects resulting from BCG immunization, while retaining cross-reactivity with the tuberculin skin test, has provided an excellent screen for those sequences that engender side effects. The identification of deletions that vary between BCG isolates identifies such sequences, which may be used in drug screening and biological analysis for the role of the deleted genes in causing untoward side effects and pathogenicity.

Identification of *M. Tuberculosis* Complex Deletion Markers

[0015] The present invention provides nucleic acid sequences that are markers for specific mycobacteria, including *M. tb.*, *M. bovis*, BCG and bacteriophage. The deletions are listed in Table 1. The absence or presence of these marker sequences is characteristic of the indicated isolate, or strain. As such, they provide a unique characteristic for the identification of the indicated mycobacteria. The deletions are identified by their *M. tb.* open reading frame ("Rv" nomenclature), which corresponds to a known genetic sequence, and may be accessed as previously cited. The junctions

of the deletions are provided by the designation of position in the publicly available *M. tb.* sequence.

Table 1

SEQ ID	rd	rv_num	orf_id	breakpoint
SEQ ID NO:1	RD01	Rv3871	MTV027.06	"H37Rv, segment 160: 7534, 16989"
SEQ ID NO:2	RD01	Rv3872	MTV027.07	"H37Rv, segment 160: 7534, 16989"
SEQ ID NO:3	RD01	Rv3873	MTV027.08	"H37Rv, segment 160: 7534, 16989"
SEQ ID NO:4	RD01	Rv3874	MTV027.09	"H37Rv, segment 160: 7534, 16989"
SEQ ID NO:5	RD01	Rv3875	MTV027.10	"H37Rv, segment 160: 7534, 16989"
SEQ ID NO:6	RD01	Rv3876	MTV027.11	"H37Rv, segment 160: 7534, 16989"
SEQ ID NO:7	RD01	Rv3877	MTV027.12	"H37Rv, segment 160: 7534, 16989"
SEQ ID NO:8	RD01	Rv3878	MTV027.13	"H37Rv, segment 160: 7534, 16989"
SEQ ID NO:9	RD01	Rv3879c	MTV027.14c	"H37Rv, segment 160: 7534, 16989"
SEQ ID NO:10	RD02	Rv1988	MTCY39.31c	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:11	RD02	Rv1987	MTCY39.32c	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:12	RD02	Rv1986	MTCY39.33c	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:13	RD02	Rv1985c	MTCY39.34	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:14	RD02	Rv1984c	MTCY39.35	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:15	RD02	Rv1983	MTCY39.36c	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:16	RD02	Rv1982c	MTCY39.37	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:17	RD02	Rv1981c	MTCY39.38	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:18	RD02	Rv1980c	MTCY39.39	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:19	RD02	Rv1979c	MTCY39.40	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:20	RD02	Rv1978	MTV051.16	"H37Rv segment 88: 14211, segment 89: 8598"
SEQ ID NO:21	RD03	Rv1586c	MTCY336.18	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:22	RD03	Rv1585c	MTCY336.19	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:23	RD03	Rv1584c	MTCY336.20	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:24	RD03	Rv1583c	MTCY336.21	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:25	RD03	Rv1582c	MTCY336.22	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:26	RD03	Rv1581c	MTCY336.23	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:27	RD03	Rv1580c	MTCY336.24	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:28	RD03	Rv1579c	MTCY336.25	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:29	RD03	Rv1578c	MTCY336.26	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:30	RD03	Rv1577c	MTCY336.27	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:31	RD03	Rv1576c	MTCY336.28	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:32	RD03	Rv1575	MTCY336.29c	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:33	RD03	Rv1574	MTCY336.30c	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:34	RD03	Rv1573	MTCY336.31c	"H37Rv, segment 70: 7677, 16923"
SEQ ID NO:35	RD04	Rv0221	MTCY08D5.16	"H37Rv, segment 12: 17432,19335"
SEQ ID NO:36	RD04	Rv0222	MTCY08D5.17	"H37Rv, segment 12: 17432,19335"
SEQ ID NO:37	RD04	Rv0223c	MTCY08D5.18	"H37Rv, segment 12: 17432,19335"
SEQ ID NO:38	RD05	Rv3117	MTCY164.27	"H37Rv, segment 135: 27437,30212"
SEQ ID NO:39	RD05	Rv3118	MTCY164.28	"H37Rv, segment 135: 27437,30212"
SEQ ID NO:40	RD05	Rv3119	MTCY164.29	"H37Rv, segment 135: 27437,30212"
SEQ ID NO:41	RD05	Rv3120	MTCY164.30	"H37Rv, segment 135: 27437,30212"
SEQ ID NO:42	RD05	Rv3121	MTCY164.31	"H37Rv, segment 135: 27437,30212"
SEQ ID NO:43	RD06	Rv1506c	MTCY277.28c	"H37Rv, segment 65: 23614, 36347"
SEQ ID NO:44	RD06	Rv1507c	MTCY277.29c	"H37Rv, segment 65: 23614, 36347"
SEQ ID NO:45	RD06	Rv1508c	MTCY277.30c	"H37Rv, segment 65: 23614, 36347"
SEQ ID NO:46	RD06	Rv1509	MTCY277.31	"H37Rv, segment 65: 23614, 36347"
SEQ ID NO:47	RD06	Rv1510	MTCY277.32	"H37Rv, segment 65: 23614, 36347"
SEQ ID NO:48	RD06	Rv1511	MTCY277.33	"H37Rv, segment 65: 23614, 36347"
SEQ ID NO:49	RD06	Rv1512	MTCY277.34	"H37Rv, segment 65: 23614, 36347"
SEQ ID NO:50	RD06	Rv1513	MTCY277.35	"H37Rv, segment 65: 23614, 36347"
SEQ ID NO:51	RD06	Rv1514c	MTCY277.36c	"H37Rv, segment 65: 23614, 36347"

SEQ ID NO:52	RD06	Rv1515c	MTCY277.37c	"H37Rv, segment 65: 23614, 36347"
SEQ ID NO:53	RD06	Rv1516c	MTCY277.38c	"H37Rv, segment 65: 23614, 36347"
SEQ ID NO:54	RD07	Rv2346c	MTCY98.15c	"H37Rv, segment 103: 17622, 26584"
SEQ ID NO:55	RD07	Rv2347c	MTCY98.16c	"H37Rv, segment 103: 17622, 26584"
SEQ ID NO:56	RD07	Rv2348c	MTCY98.17c	"H37Rv, segment 103: 17622, 26584"
SEQ ID NO:57	RD07	Rv2349c	MTCY98.18c	"H37Rv, segment 103: 17622, 26584"
SEQ ID NO:58	RD07	Rv2350c	MTCY98.19c	"H37Rv, segment 103: 17622, 26584"
SEQ ID NO:59	RD07	Rv2351c	MTCY98.20c	"H37Rv, segment 103: 17622, 26584"
SEQ ID NO:60	RD07	Rv2352c	MTCY98.21c	"H37Rv, segment 103: 17622, 26584"
SEQ ID NO:61	RD07	Rv2353c	MTCY98.22c	"H37Rv, segment 103: 17622, 26584"
SEQ ID NO:62	RD08	Rv0309	MTCY63.14	"H37Rv, segment 16: 17018, 20446"
SEQ ID NO:63	RD08	Rv0310c	MTCY63.15c	"H37Rv, segment 16: 17018, 20446"
SEQ ID NO:64	RD08	Rv0311	MTCY63.16	"H37Rv, segment 16: 17018, 20446"
SEQ ID NO:65	RD08	Rv0312	MTCY63.17	"H37Rv, segment 16: 17018, 20446"
SEQ ID NO:66	RD09	Rv3623	MTCY15C10.29c	"H37Rv, segment 153: 21131, segment 154: 2832"
SEQ ID NO:67	RD09	Rv3622c	MTCY15C10.30	"H37Rv, segment 153: 21131, segment 154: 2832"
SEQ ID NO:68	RD09	Rv3621c	MTCY15C10.31	"H37Rv, segment 153: 21131, segment 154: 2832"
SEQ ID NO:69	RD09	Rv3620c	MTCY15C10.32	"H37Rv, segment 153: 21131, segment 154: 2832"
SEQ ID NO:70	RD09	Rv3619c	MTCY15C10.33	"H37Rv, segment 153: 21131, segment 154: 2832"
SEQ ID NO:71	RD09	Rv3618	MTCY15C10.34c	"H37Rv, segment 153: 21131, segment 154: 2832"
SEQ ID NO:72	RD09	Rv3617	MTCY15C10.35c	"H37Rv, segment 153: 21131, segment 154: 2832"
SEQ ID NO:73	RD10	Rv1257c	MTCY50.25	"H37Rv segment 55: 3689, 6696"
SEQ ID NO:74	RD10	Rv1256c	MTCY50.26	"H37Rv segment 55: 3689, 6696"
SEQ ID NO:75	RD10	Rv1255c	MTCY50.27	"H37Rv segment 55: 3689, 6696"
SEQ ID NO:76	RD11	Rv3429	MTCY77.01	"H37Rv, segment 145: 30303 to segment 146: 1475"
SEQ ID NO:77	RD11	Rv3428c	MTCY78.01	"H37Rv, segment 145: 30303 to segment 146: 1475"
SEQ ID NO:78	RD11	Rv3427c	MTCY78.02	"H37Rv, segment 145: 30303 to segment 146: 1475"
SEQ ID NO:79	RD11	Rv3426	MTCY78.03c	"H37Rv, segment 145: 30303 to segment 146: 1475"
SEQ ID NO:80	RD11	Rv3425	MTCY78.04c	"H37Rv, segment 145: 30303 to segment 146: 1475"
SEQ ID NO:81	RD12	Rv2072c	MTCY49.11c	"H37Rv segment 93: 9301, 11331"
SEQ ID NO:82	RD12	Rv2073c	MTCY49.12c	"H37Rv segment 93: 9301, 11331"
SEQ ID NO:83	RD12	Rv2074	MTCY49.13	"H37Rv segment 93: 9301, 11331"
SEQ ID NO:84	RD12	Rv2075c	MTCY49.14c	"H37Rv segment 93: 9301, 11331"
SEQ ID NO:85	RD13bis	Rv2645	MTCY441.15	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:86	RD13bis	Rv2646	MTCY441.16	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:87	RD13bis	Rv2647	MTCY441.17	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:88	RD13bis	Rv2648	MTCY441.17A	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:89	RD13bis	Rv2649	MTCY441.18	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:90	RD13bis	Rv2650c	MTCY441.19	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:91	RD13bis	Rv2651c	MTCY441.20c	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:92	RD13bis	Rv2652c	MTCY441.21c	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:93	RD13bis	Rv2653c	MTCY441.22c	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:94	RD13bis	Rv2654c	MTCY441.23c	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:95	RD13bis	Rv2655c	MTCY441.24c	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:96	RD13bis	Rv2656c	MTCY441.25c	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:97	RD13bis	Rv2657c	MTCY441.26c	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:98	RD13bis	Rv2658c	MTCY441.27c	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:99	RD13bis	Rv2659c	MTCY441.28c	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:100	RD13bis	Rv2660c	MTCY441.29c	"H37Rv, segment 118: 12475, 23455"
SEQ ID NO:101	RD14	Rv1766	MTCY28.32	"H37Rv segment 79: 30573, 39642"
SEQ ID NO:102	RD14	Rv1767	MTCY28.33	"H37Rv segment 79: 30573, 39642"
SEQ ID NO:103	RD14	Rv1768	MTCY28.34	"H37Rv segment 79: 30573, 39642"
SEQ ID NO:104	RD14	Rv1769	MTCY28.35	"H37Rv segment 79: 30573, 39642"
SEQ ID NO:105	RD14	Rv1770	MTCY28.36	"H37Rv segment 79: 30573, 39642"
SEQ ID NO:106	RD14	Rv1771	MTCY28.37	"H37Rv segment 79: 30573, 39642"
SEQ ID NO:107	RD14	Rv1772	MTCY28.38	"H37Rv segment 79: 30573, 39642"
SEQ ID NO:108	RD14	Rv1773c	MTCY28.39	"H37Rv segment 79: 30573, 39642"
SEQ ID NO:109	RD15	Rv1963c	MTV051.01c	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:110	RD15	Rv1964	MTV051.02	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:111	RD15	Rv1965	MTV051.03	"H37Rv segment 88: 1153, 13873"

SEQ ID NO:112	RD15	Rv1966	MTV051.04	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:113	RD15	Rv1967	MTV051.05	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:114	RD15	Rv1968	MTV051.06	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:115	RD15	Rv1969	MTV051.07	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:116	RD15	Rv1970	MTV051.08	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:117	RD15	Rv1971	MTV051.09	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:118	RD15	Rv1972	MTV051.10	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:119	RD15	Rv1973	MTV051.11	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:120	RD15	Rv1974	MTV051.12	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:121	RD15	Rv1975	MTV051.13	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:122	RD15	Rv1976c	MTV051.14	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:123	RD15	Rv1977	MTV051.15	"H37Rv segment 88: 1153, 13873"
SEQ ID NO:124	RD16	Rv3405c	MTCY78.23	"H37Rv, segment 145: 5012, 12621"
SEQ ID NO:125	RD16	Rv3404c	MTCY78.24	"H37Rv, segment 145: 5012, 12621"
SEQ ID NO:126	RD16	Rv3403c	MTCY78.25	"H37Rv, segment 145: 5012, 12621"
SEQ ID NO:127	RD16	Rv3402c	MTCY78.26	"H37Rv, segment 145: 5012, 12621"
SEQ ID NO:128	RD16	Rv3401	MTCY78.27c	"H37Rv, segment 145: 5012, 12621"
SEQ ID NO:129	RD16	Rv3400	MTCY78.28c	"H37Rv, segment 145: 5012, 12621"

[0016] The "Rv" column indicates public *M. tb* sequence, open reading frame. The BCG strains were obtained as follows:

Table 2. Strains employed in study of BCG phylogeny

Name of strain	Synonym	Source	Descriptors
BCG-Russia	Moscow	ATCC	# 35740
BCG-Moreau	Brazil	ATCC	# 35736
BCG-Moreau	Brazil	IAF	dated 1958
BCG-Moreau	Brazil	IAF	dated 1961
BCG-Japan	Tokyo	ATCC	# 35737
BCG-Japan	Tokyo	IAF	dated 1961
BCG-Japan	Tokyo	JATA	vaccine strain
BCG-Japan	Tokyo	JATA	bladder cancer strain
BCG-Japan	Tokyo	JATA	clinical isolate- adenitis
BCG-Sweden	Gothenburg	ATCC	# 35732
BCG-Sweden	Gothenburg	IAF	dated 1958
BCG-Sweden	Gothenburg	SSI	production lot, Copenhagen
BCG-Phipps	Philadelphia	ATCC	# 35744
BCG-Denmark	Danish 1331	ATCC	# 35733
BCG-Copenhagen		ATCC	#27290
BCG-Copenhagen		IAF	dated 1961
BCG-Tice	Chicago	vaccine	dated 1973
BCG-Tice	Chicago	ATCC	# 35743
BCG-Frappier	Montreal	IAF	primary lot, 1973
BCG-Frappier, INH-resistant	Montreal-R	IAF	primary lot, 1973
BCG-Frappier	Montreal	IAF	passage 946
BCG-Connaught	Toronto	CL	bladder cancer treatment
BCG-Birkhaug		ATCC	# 35731

BCG-Prague	Czech	SSI	lyophilized 1968
BCG-Glaxo		vaccine	dated 1973
BCG-Glaxo		ATCC	# 35741
BCG-Pasteur		IAF	passage 888
BCG-Pasteur		IAF	dated 1961
BCG-Pasteur		IP	1173P2-B
BCG-Pasteur		IP	1173P2-C
BCG-Pasteur		IP	clinical isolate # 1
BCG-Pasteur		IP	clinical isolate # 2
BCG-Pasteur		ATCC	# 35734

Abbreviations: IP= Institut Pasteur, Paris, France; IAF= Institut Armand Frappier, Laval, Canada; ATCC= American Type Culture Collection, Rockville, Md, USA; SSI=Statens Serum Institute, Copenhagen, Denmark; CL=Connaught Laboratories, Willowdale, Canada, JATA= Japanese Anti-Tuberculosis Association; INH =isoniazid. Canadian BCG's refers to BCG-Montreal and BCG-Toronto, the latter being derived from the former.

[0017] In performing the initial screening method, genomic DNA is isolated from two mycobacteria microbial cell cultures. The two DNA preparations are labeled, where a different label is used for the first and second microbial cultures, typically using nucleotides conjugated to a fluorochrome that emits at a wavelength substantially different from that of the fluorochrome tagged nucleotides used to label the selected probe. The strains used were the reference strain of *Mycobacterium tuberculosis* (H37Rv), other *M. tb.* laboratory strains, such as H37Ra, the O strain, *M. tb.* clinical isolates, the reference strain of *Mycobacterium bovis*, and different strains of *Mycobacterium bovis* BCG.

[0018] The two DNA preparations are mixed, and competitive hybridization is carried out to a microarray representing all of the open reading frames in the genome of the test microbe, usually H37Rv. Hybridization of the labeled sequences is accomplished according to methods well known in the art. In a preferred embodiment, the two probes are combined to provide for a competitive hybridization to a single microarray. Hybridization can be carried out under conditions varying in stringency, preferably under conditions of high stringency (*e.g.*, 4x SSC, 10% SDS, 65° C) to allow for hybridization of complementary sequences having extensive homology (*e.g.*, having at least 85% sequence identity, preferably at least 90% sequence identity, more preferably having at least 95% sequence identity). Where the target sequences are native sequences the hybridization is preferably carried out under

conditions that allow hybridization of only highly homologous sequences (*e.g.*, at least 95% to 100% sequence identity).

[0019] Two color fluorescent hybridization is utilized to assay the representation of the unselected library in relation to the selected library (*i.e.*, to detect hybridization of the unselected probe relative to the selected probe). From the ratio of one color to the other, for any particular array element, the relative abundance of that sequence in the unselected and selected libraries can be determined. In addition, comparison of the hybridization of the selected and unselected probes provides an internal control for the assay. An absence of signal from the reference strain, as compared to H37Rv, is indicative that the open reading frame is deleted in the test strain. The deletion may be further mapped by Southern blot analysis, and by sequencing the regions flanking the deletion.

[0020] Microarrays can be scanned to detect hybridization of the selected and the unselected sequences using a custom built scanning laser microscope as described in Shalon *et al.*, *Genome Res.* 6:639 (1996). A separate scan, using the appropriate excitation line, is performed for each of the two fluorophores used. The digital images generated from the scan are then combined for subsequent analysis. For any particular array element, the ratio of the fluorescent signal from the amplified selected cell population DNA is compared to the fluorescent signal from the unselected cell population DNA, and the relative abundance of that sequence in the selected and unselected library determined.

Nucleic Acid Compositions

[0021] As used herein, the term "deletion marker", or "marker" is used to refer to those sequences of *M. tuberculosis* complex genomes that are deleted in one or more of the strains or species, as indicated in Table 1. The bacteria of the *M. tuberculosis* complex include *M. tuberculosis*, *M. bovis*, and BCG, inclusive of varied isolates and strains within each species. Nucleic acids of interest include all or a portion of the deleted region, particularly complete open reading frames, hybridization primers, promoter regions, *etc.*

[0022] The term "junction" or "deletion junction" is used to refer to nucleic acids that comprise the regions on both the 3' and the 5' sequence immediately flanking the deletion. Such junction sequences are preferably used as short primers, *e.g.* from about 15 nt to about 30 nt, that specifically hybridize to the junction, but not to a nucleic acid comprising the undeleted genomic sequence. For example, the deletion found in M. bovis, at Rv0221, corresponds to the nucleotide sequence of the M. tuberculosis H37Rv genome, segment 12: 17432,19335. The junction comprises the regions upstream of position 17342, and downstream of 19335, *e.g.* a nucleic acid of 20 nucleotides comprising the sequence from H37Rv 17332-17342 joined to 19335-19345.

[0023] Typically, such nucleic acids comprising a junction will include at least about 7 nucleotides from each flanking region, *i.e.* from the 3' and from the 5' sequences adjacent to the deletion, and may be about 10 nucleotides from each flanking region, up to about 15 nucleotides, or more. Amplification primers that hybridize to the junction sequence, to the deleted sequence, and to the flanking non-deleted regions have a variety of uses, as detailed below.

[0024] The nucleic acid compositions of the subject invention encode all or a part of the deletion markers. Fragments may be obtained of the DNA sequence by chemically synthesizing oligonucleotides in accordance with conventional methods, by restriction enzyme digestion, by PCR amplification, *etc.* For the most part, DNA fragments will be at least about 25 nt in length, usually at least about 30 nt, more usually at least about 50 nt. For use in amplification reactions, such as PCR, a pair of primers will be used. The exact composition of the primer sequences is not critical to the invention, but for most applications the primers will hybridize to the subject sequence under stringent conditions, as known in the art. It is preferable to chose a pair of primers that will generate an amplification product of at least about 50 nt, preferably at least about 100 nt. Algorithms for the selection of primer sequences are generally known, and are available in commercial software packages. Amplification primers hybridize to complementary strands of DNA, and will prime towards each other.

[0025] Usually, the DNA will be obtained substantially free of other nucleic acid sequences that do not include a deletion marker sequence or fragment thereof, generally being at least about 50%, usually at least about 90% pure and are typically "recombinant", *i.e.* flanked by one or more nucleotides with which it is not normally associated on a naturally occurring chromosome.

[0026] For screening purposes, hybridization probes of one or more of the deletion sequences may be used in separate reactions or spatially separated on a solid phase matrix, or labeled such that they can be distinguished from each other. Assays may utilize nucleic acids that hybridize to one or more of the described deletions.

[0027] An array may include all or a subset of the deletion markers listed in Table 1. Usually such an array will include at least 2 different deletion marker sequences, *i.e.* deletions located at unique positions within the locus, and may include all of the provided deletion markers. Arrays of interest may further comprise other genetic sequences, particularly other sequences of interest for tuberculosis screening. The oligonucleotide sequence on the array will usually be at least about 12 nt in length, may be the length of the provided deletion marker sequences, or may extend into the flanking regions to generate fragments of 100 to 200 nt in length. For examples of arrays, see Ramsay (1998) Nat. Biotech. **16**:40-44; Hacia *et al.* (1996) Nature Genetics **14**:441-447; Lockhart *et al.* (1996) Nature Biotechnol. **14**:1675-1680; and De Risi *et al.* (1996) Nature Genetics **14**:457-460.

[0028] Nucleic acids may be naturally occurring, *e.g.* DNA or RNA, or may be synthetic analogs, as known in the art. Such analogs may be preferred for use as probes because of superior stability under assay conditions. Modifications in the native structure, including alterations in the backbone, sugars or heterocyclic bases, have been shown to increase intracellular stability and binding affinity. Among useful changes in the backbone chemistry are phosphorothioates; phosphorodithioates, where both of the non-bridging oxygens are substituted with sulfur; phosphoroamidites; alkyl phosphotriesters and boranophosphates. Achiral phosphate derivatives include 3'-O'-5'-S-phosphorothioate, 3'-S-5'-O-phosphorothioate, 3'-CH_2-5'-O-phosphonate and 3'-NH-5'-O-phosphoroamidate.

Peptide nucleic acids replace the entire ribose phosphodiester backbone with a peptide linkage.

[0029] Sugar modifications are also used to enhance stability and affinity. The α-anomer of deoxyribose may be used, where the base is inverted with respect to the natural b-anomer. The 2'-OH of the ribose sugar may be altered to form 2'-O-methyl or 2'-O-allyl sugars, which provide resistance to degradation without comprising affinity.

[0030] Modification of the heterocyclic bases must maintain proper base pairing. Some useful substitutions include deoxyuridine for deoxythymidine; 5-methyl-2'-deoxycytidine and 5-bromo-2'-deoxycytidine for deoxycytidine. 5- propynyl-2'-deoxyuridine and 5-propynyl-2'-deoxycytidine have been shown to increase affinity and biological activity when substituted for deoxythymidine and deoxycytidine, respectively.

Polypeptide Compositions

[0031] The specific deletion markers in Table 1 correspond to open reading frames of the *M. tb* genome, and therefore encode a polypeptide. The subject markers may be employed for synthesis of a complete protein, or polypeptide fragments thereof, particularly fragments corresponding to functional domains; binding sites; *etc.*; and including fusions of the subject polypeptides to other proteins or parts thereof. For expression, an expression cassette may be employed, providing for a transcriptional and translational initiation region, which may be inducible or constitutive, where the coding region is operably linked under the transcriptional control of the transcriptional initiation region, and a transcriptional and translational termination region. Various transcriptional initiation regions may be employed that are functional in the expression host.

[0032] The polypeptides may be expressed in prokaryotes or eukaryotes in accordance with conventional ways, depending upon the purpose for expression. For large scale production of the protein, a unicellular organism, such as *E. coli, B.*

subtilis, S. cerevisiae, or cells of a higher organism such as vertebrates, particularly mammals, *e.g.* COS 7 cells, may be used as the expression host cells. Small peptides can also be synthesized in the laboratory.

[0033] With the availability of the polypeptides in large amounts, by employing an expression host, the polypeptides may be isolated and purified in accordance with conventional ways. A lysate may be prepared of the expression host and the lysate purified using HPLC, exclusion chromatography, gel electrophoresis, affinity chromatography, or other purification technique. The purified polypeptide will generally be at least about 80% pure, preferably at least about 90% pure, and may be up to and including 100% pure. Pure is intended to mean free of other proteins, as well as cellular debris.

[0034] The polypeptide is used for the production of antibodies, where short fragments provide for antibodies specific for the particular polypeptide, and larger fragments or the entire protein allow for the production of antibodies over the surface of the polypeptide. Antibodies may be raised to isolated peptides corresponding to particular domains, or to the native protein.

[0035] Antibodies are prepared in accordance with conventional ways, where the expressed polypeptide or protein is used as an immunogen, by itself or conjugated to known immunogenic carriers, *e.g.* KLH, pre-S HBsAg, other viral or eukaryotic proteins, or the like. Various adjuvants may be employed, with a series of injections, as appropriate. For monoclonal antibodies, after one or more booster injections, the spleen is isolated, the lymphocytes immortalized by cell fusion, and then screened for high affinity antibody binding. The immortalized cells, *i.e.* hybridomas, producing the desired antibodies may then be expanded. For further description, see Monoclonal Antibodies: A Laboratory Manual, Harlow and Lane eds., Cold Spring Harbor Laboratories, Cold Spring Harbor, New York, 1988. If desired, the mRNA encoding the heavy and light chains may be isolated and mutagenized by cloning in *E. coli*, and the heavy and light chains mixed to further enhance the affinity of the antibody. Alternatives to *in vivo* immunization as a method of raising antibodies

include binding to phage "display" libraries, usually in conjunction with *in vitro* affinity maturation.

[0036] The antibody may be produced as a single chain, instead of the normal multimeric structure. Single chain antibodies are described in Jost *et al.* (1994) J.B.C. **269**:26267–73, and others. DNA sequences encoding the variable region of the heavy chain and the variable region of the light chain are ligated to a spacer encoding at least about 4 amino acids of small neutral amino acids, including glycine and/or serine. The protein encoded by this fusion allows assembly of a functional variable region that retains the specificity and affinity of the original antibody.

Use of Deletion Markers in Identification of Mycobacteria

[0037] The deletions provided in Table 1 are useful for the identification of a mycobacterium as (a) variants of *M. tb.* (b) isolates of BCG (c) *M. bovis* strains or (d) carrying the identified mycobacterial bacteriophage, depending on the specific marker that is chosen. Such screening is particularly useful in determining whether a particular infection or isolate is pathogenic. The term mycobacteria may refer to any member of the family Mycobacteriacaeae, including *M. tuberculosis*, *M. avium complex*, *M. kansasii*, *M. scrofulaceum*, *M. bovis* and *M. leprae*.

[0038] Means of detecting deletions are known in the art. Deletions may be identified through the absence or presence of the sequences in mRNA or genomic DNA, through analysis of junctional regions that flank the deletion, or detection of the gene product, or, particularly relating to the tuberculin skin test, by identification of antibodies that react with the encoded gene product.

[0039] While deletions can be easily determined by the absence of hybridization, in many cases it is desirable to have a positive signal, in order to minimize artifactual negative readings. In such cases the deletions may be detected by designing a primer that flanks the junction formed by the deletion. Where the deletion is present, a novel sequence is formed between the flanking regions, which can be detected by hybridization. Preferably such a primer will be sufficiently short that it will only

hybridize to the junction, and will fail to form stable hybrids with either of the separate parts of the junction.

[0040] Diagnosis is performed by protein, DNA or RNA sequence and/or hybridization analysis of any convenient sample, *e.g.* cultured mycobacteria, biopsy material, blood sample, *etc.* Screening may also be based on the functional or antigenic characteristics of the protein. Immunoassays designed to detect the encoded proteins from deleted sequences may be used in screening.

[0041] A number of methods are available for analyzing nucleic acids for the presence of a specific sequence. Where large amounts of DNA are available, genomic DNA is used directly. Alternatively, the region of interest is cloned into a suitable vector and grown in sufficient quantity for analysis. The nucleic acid may be amplified by conventional techniques, such as the polymerase chain reaction (PCR), to provide sufficient amounts for analysis. The use of the polymerase chain reaction is described in Saiki, *et al.* (1985) Science **239**:487, and a review of current techniques may be found in Sambrook, *et al.* Molecular Cloning: A Laboratory Manual, CSH Press 1989, pp.14.2-14.33. Amplification may also be used to determine whether a polymorphism is present, by using a primer that is specific for the polymorphism. Alternatively, various methods are known in the art that utilize oligonucleotide ligation, for examples see Riley *et al.* (1990) N.A.R. **18**:2887-2890; and Delahunty *et al.* (1996) Am. J. Hum. Genet. **58**:1239-1246.

[0042] A detectable label may be included in an amplification reaction. Suitable labels include fluorochromes, *e.g.* fluorescein isothiocyanate (FITC), rhodamine, Texas Red, phycoerythrin, allophycocyanin, 6-carboxyfluorescein (6-FAM), 2',7'-dimethoxy-4',5'-dichloro-6-carboxyfluorescein (JOE), 6-carboxy-X-rhodamine (ROX), 6-carboxy-2',4',7',4,7-hexachlorofluorescein (HEX), 5-carboxyfluorescein (5-FAM) or N,N,N',N'-tetramethyl-6-carboxyrhodamine (TAMRA), radioactive labels, *e.g.* ^{32}P, ^{35}S, ^{3}H; *etc.* The label may be a two stage system, where the amplified DNA is conjugated to biotin, haptens, *etc.* having a high affinity binding partner, *e.g.* avidin, specific antibodies, *etc.*, where the binding partner is conjugated to a detectable label. The label may be conjugated to one or both of the primers.

Alternatively, the pool of nucleotides used in the amplification is labeled, so as to incorporate the label into the amplification product.

[0043] The sample nucleic acid, *e.g.* amplified or cloned fragment, is analyzed by one of a number of methods known in the art. The nucleic acid may be sequenced by dideoxy or other methods, and the sequence of bases compared to the deleted sequence. Hybridization with the variant sequence may also be used to determine its presence, by Southern blots, dot blots, *etc.* The hybridization pattern of a control and variant sequence to an array of oligonucleotide probes immobilized on a solid support, as described in US 5,445,934, or in WO95/35505, may also be used as a means of detecting the presence of variable sequences. Single strand conformational polymorphism (SSCP) analysis, denaturing gradient gel electrophoresis (DGGE), mismatch cleavage detection, and heteroduplex analysis in gel matrices are used to detect conformational changes created by DNA sequence variation as alterations in electrophoretic mobility. Alternatively, where a polymorphism creates or destroys a recognition site for a restriction endonuclease (restriction fragment length polymorphism, RFLP), the sample is digested with that endonuclease, and the products size fractionated to determine whether the fragment was digested. Fractionation is performed by gel or capillary electrophoresis, particularly acrylamide or agarose gels.

[0044] The hybridization pattern of a control and variant sequence to an array of oligonucleotide probes immobilized on a solid support, as described in US 5,445,934, or in WO95/35505, may be used as a means of detecting the presence or absence of deleted sequences. In one embodiment of the invention, an array of oligonucleotides is provided, where discrete positions on the array are complementary to at least a portion of M. tb. genomic DNA, usually comprising at least a portion from the identified open reading frames. Such an array may comprise a series of oligonucleotides, each of which can specifically hybridize to a nucleic acid, *e.g.* mRNA, cDNA, genomic DNA, *etc.*

[0045] Deletions may also be detected by amplification. In an embodiment of the invention, sequences are amplified that include a deletion junction, *i.e.* where the

amplification primers hybridize to a junction sequence. In a nucleic acid sample where the marker sequence is deleted, a junction will be formed, and the primer will hybridize, thereby allowing amplification of a detectable sequence. In a nucleic acid sample where the marker sequence is present, the primer will not hybridize, and no amplification will take place. Alternatively, amplification primers may be chosen such that amplification of the target sequence will only take place where the marker sequence is present. The amplification products may be separated by size using any convenient method, as known in the art, including gel electrophoresis, chromatography, capillary electrophoresis, density gradient fractionation, *etc.*

[0046] In addition to the detection of deletions by the detection of junctions sequences, or detection of the marker sequences themselves, one may determine the presence or absence of the encoded protein product. The specific deletions in Table 1 correspond to open reading frames of the *M. tb* genome, and therefore encode polypeptides. Polypeptides are detected by means known in the art, including determining the presence of the specific polypeptide in a sample through biochemical, functional or immunological characterization. The detection of antibodies in patient serum that react with a polypeptide is of particular interest.

[0047] Immunization with BCG typically leads to a positive response against tuberculin antigens in a skin test. In people who have been immunized, which includes a significant proportion of the world population, it is therefore difficult to determine whether a positive test is the result of an immune reaction to the BCG vaccine, or to an ongoing *M. tb*. infection. The subject invention has provided a number of open reading frame sequences that are present in M. tb isolates, but are absent in BCG. As a primary or a secondary screening method, one may test for immunoreactivity of the patient with the polypeptides encoded by such deletion markers. Diagnosis may be performed by a number of methods. The different methods all determine the presence of an immune response to the polypeptide in a patient, where a positive response is indicative of an *M. tb* infection. The immune response may be determined by determination of antibody binding, or by the presence of a response to intradermal challenge with the polypeptide.

[0048] In one method, a dose of the deletion marker polypeptide, formulated as a cocktail of proteins or as individual protein species, in a suitable medium is injected subcutaneously into the patient. The dose will usually be at least about 0.05 μg of protein, and usually not more than about 5 μg of protein. A control comprising medium alone, or an unrelated protein will be injected nearby at the same time. The site of injection is examined after a period of time for the presence of a wheal. The wheal at the site of polypeptide injection is compared to that at the site of the control injection, usually by measuring the size of the wheal. The skin test readings may be assessed by a variety of objective grading systems. A positive result for the presence of an allergic condition will show an increased diameter at the site of polypeptide injection as compared to the control, usually at least about 50% increase in size, more usually at least 100% increase in size.

[0049] An alternative method for diagnosis depends on the *in vitro* detection of binding between antibodies in a patient sample and the subject polypeptides, either as a cocktail or as individual protein species, where the presence of specific binding is indicative of an infection. Measuring the concentration of polypeptide specific antibodies in a sample or fraction thereof may be accomplished by a variety of specific assays. In general, the assay will measure the reactivity between a patient sample, usually blood derived, generally in the form of plasma or serum. The patient sample may be used directly, or diluted as appropriate, usually about 1:10 and usually not more than about 1:10,000. Immunoassays may be performed in any physiological buffer, *e.g.* PBS, normal saline, HBSS, dPBS, *etc.*

[0050] In a preferred embodiment, a conventional sandwich type assay is used. A sandwich assay is performed by first attaching the polypeptide to an insoluble surface or support. The polypeptide may be bound to the surface by any convenient means, depending upon the nature of the surface, either directly or through specific antibodies. The particular manner of binding is not crucial so long as it is compatible with the reagents and overall methods of the invention. They may be bound to the plates covalently or non-covalently, preferably non-covalently. Samples, fractions or aliquots thereof are then added to separately assayable supports (for example, separate

wells of a microtiter plate) containing support-bound polypeptide. Preferably, a series of standards, containing known concentrations of antibodies is assayed in parallel with the samples or aliquots thereof to serve as controls.

[0051] Immune specific receptors may be labeled to facilitate direct, or indirect quantification of binding. Examples of labels which permit direct measurement of second receptor binding include radiolabels, such as ^{3}H or ^{125}I, fluorescers, dyes, beads, chemilumninescers, colloidal particles, and the like. Examples of labels which permit indirect measurement of binding include enzymes where the substrate may provide for a colored or fluorescent product. In a preferred embodiment, the second receptors are antibodies labeled with a covalently bound enzyme capable of providing a detectable product signal after addition of suitable substrate. Examples of suitable enzymes for use in conjugates include horseradish peroxidase, alkaline phosphatase, malate dehydrogenase and the like. Where not commercially available, such antibody-enzyme conjugates are readily produced by techniques known to those skilled in the art.

[0052] In some cases, a competitive assay will be used. In addition to the patient sample, a competitor to the antibody is added to the reaction mix. The competitor and the antibody compete for binding to the polypeptide. Usually, the competitor molecule will be labeled and detected as previously described, where the amount of competitor binding will be proportional to the amount of Immune present. The concentration of competitor molecule will be from about 10 times the maximum anticipated Immune concentration to about equal concentration in order to make the most sensitive and linear range of detection.

[0053] Alternatively, antibodies may be used for direct determination of the presence of the deletion marker polypeptide. Antibodies specific for the subject deletion markers as previously described may be used in screening immunoassays. Samples, as used herein, include microbial cultures, biological fluids such as tracheal lavage, blood, etc. Also included in the term are derivatives and fractions of such fluids. Diagnosis may be performed by a number of methods. The different methods all determine the absence or presence of polypeptides encoded by the subject deletion

markers. For example, detection may utilize staining of mycobacterial cells or histological sections, performed in accordance with conventional methods. The antibodies of interest are added to the cell sample, and incubated for a period of time sufficient to allow binding to the epitope, usually at least about 10 minutes. The antibody may be labeled with radioisotopes, enzymes, fluorescers, chemiluminescers, or other labels for direct detection. Alternatively, a second stage antibody or reagent is used to amplify the signal. Such reagents are well known in the art. For example, the primary antibody may be conjugated to biotin, with horseradish peroxidase-conjugated avidin added as a second stage reagent. Final detection uses a substrate that undergoes a color change in the presence of the peroxidase. The absence or presence of antibody binding may be determined by various methods, including microscopy, radiography, scintillation counting, *etc.*

[0054] An alternative method for diagnosis depends on the *in vitro* detection of binding between antibodies and the subject polypeptides in solution, e.g. a cell lysate. Measuring the concentration of binding in a sample or fraction thereof may be accomplished by a variety of specific assays. A conventional sandwich type assay may be used. For example, a sandwich assay may first attach specific antibodies to an insoluble surface or support. The particular manner of binding is not crucial so long as it is compatible with the reagents and overall methods of the invention. They may be bound to the plates covalently or non-covalently, preferably non-covalently. The insoluble supports may be any compositions to which polypeptides can be bound, which is readily separated from soluble material, and which is otherwise compatible with the overall method. The surface of such supports may be solid or porous and of any convenient shape. Examples of suitable insoluble supports to which the receptor is bound include beads, *e.g.* magnetic beads, membranes and microtiter plates. These are typically made of glass, plastic (*e.g.* polystyrene), polysaccharides, nylon or nitrocellulose. Microtiter plates are especially convenient because a large number of assays can be carried out simultaneously, using small amounts of reagents and samples.

[0055] Samples are then added to separately assayable supports (for example, separate wells of a microtiter plate) containing antibodies. Preferably, a series of standards, containing known concentrations of the polypeptides is assayed in parallel with the samples or aliquots thereof to serve as controls. Preferably, each sample and standard will be added to multiple wells so that mean values can be obtained for each. The incubation time should be sufficient for binding, generally, from about 0.1 to 3 hr is sufficient. After incubation, the insoluble support is generally washed of non-bound components. Generally, a dilute non-ionic detergent medium at an appropriate pH, generally 7-8, is used as a wash medium. From one to six washes may be employed, with sufficient volume to thoroughly wash non-specifically bound proteins present in the sample.

[0056] After washing, a solution containing a second antibody is applied. The antibody will bind with sufficient specificity such that it can be distinguished from other components present. The second antibodies may be labeled to facilitate direct, or indirect quantification of binding. Examples of labels that permit direct measurement of second receptor binding include radiolabels, such as ^{3}H or ^{125}I, fluorescers, dyes, beads, chemilumninescers, colloidal particles, and the like. Examples of labels which permit indirect measurement of binding include enzymes where the substrate may provide for a colored or fluorescent product. In a preferred embodiment, the antibodies are labeled with a covalently bound enzyme capable of providing a detectable product signal after addition of suitable substrate. Examples of suitable enzymes for use in conjugates include horseradish peroxidase, alkaline phosphatase, malate dehydrogenase and the like. Where not commercially available, such antibody-enzyme conjugates are readily produced by techniques known to those skilled in the art. The incubation time should be sufficient for the labeled ligand to bind available molecules. Generally, from about 0.1 to 3 hr is sufficient, usually 1 hr sufficing.

[0057] After the second binding step, the insoluble support is again washed free of non-specifically bound material. The signal produced by the bound conjugate is

detected by conventional means. Where an enzyme conjugate is used, an appropriate enzyme substrate is provided so a detectable product is formed.

[0058] Other immunoassays are known in the art and may find use as diagnostics. Ouchterlony plates provide a simple determination of antibody binding. Western blots may be performed on protein gels or protein spots on filters, using a detection system specific for the polypeptide, conveniently using a labeling method as described for the sandwich assay.

Recombinant Mycobacterium

[0059] Mycobacterium, particularly those of the *M. tuberculosis* complex, are genetically engineered to contain specific deletions or insertions corresponding to the identified genetic markers. In particular, attenuated BCG strains are modified to introduce deleted genes encoding sequences important in the establishment of effective immunity. Alternatively, *M. bovis* or *M. tuberculosis* are modified by homologous recombination to create specific deletions in sequences that determine virulence, *i.e.* the bacteria are attenuated through recombinant techniques.

[0060] In order to stably introduce sequences into BCG, the *M. tb* open reading frame corresponding to one of the deletions in Table 1 is inserted into a vector that is maintained in *M. bovis* strains. Preferably, the native 5' and 3' flanking sequences are included, in order to provide for suitable regulation of transcription and translation. However, in special circumstances, exogenous promoters and other regulatory regions may be included. Vectors and methods of transfection for BCG are known in the art. For example, U.S. Patent no. 5,776,465, herein incorporated by reference, describes the introduction of exogenous genes into BCG.

[0061] In one embodiment of the invention, the complete deleted region is replaced in BCG. The junctions of the deletion are determined as compared to a wild type *M. tb.* or *M. bovis* sequence, for example as set forth in the experimental section. The deleted region is cloned by any convenient method, as known in the art, *e.g.* PCR amplification of the region, restriction endonuclease digestion, chemical synthesis, *etc.* Preferably the cloned region will further comprise flanking sequences of a length

sufficient to induce homologous recombination, usually at least about 25 nt, more usually at least about 100 nt, or greater. Suitable vectors and methods are known in the art, for an example, see Norman *et al.* (1995) Mol. Microbiol. **16**:755-760.

[0062] In an alternative embodiment, one or more of the deletions provided in Table 1 are introduced into a strain of *M. tuberculosis* or *M. bovis*. Preferably such a strain is reduced in virulence, *e.g.* H37Ra, *etc.* Methods of homologous recombination in order to effect deletions in mycobacteria are known in the art, for example, see Norman *et al.*, *supra.*; Ganjam *et al.* (1991) P.N.A.S. **88**:5433-5437; and Aldovini *et al.* (1993) J. Bacteriol. **175**:7282-7289. Deletions may comprise an open reading frame identified in Table 1, or may extend to the full deletion, *i.e.* extending into flanking regions, and may include multiple open reading frames.

[0063] The ability of the genetically altered mycobacterium to cause disease may be tested in one or more experimental models. For example, *M. tb.* is known to infect a variety of animals, and cells in culture. In one assay, mammalian macrophages, preferably human macrophages, are infected. In a comparison of virulent, avirulent and attenuated strains of the *M. tuberculosis* complex, alveolar or peripheral blood monocytes are infected at a 1:1 ratio (Silver *et al.* (1998) Infect Immun **66**(3):1190-1199; Paul *et al.* (1996) J Infect Dis **174**(1):105-112.) The percentages of cells infected by the strains and the initial numbers of intracellular organisms are equivalent, as were levels of monocyte viability up to 7 days following infection. However, intracellular growth reflects virulence, over a period of one or more weeks. Mycobacterial growth may be evaluated by acid-fast staining, electron microscopy, and colony-forming units (cfu) assays. Monocyte production of tumor necrosis factor alpha may also be monitored as a marker for virulence.

[0064] Other assays for virulence utilize animal models. The *M. tb.* complex bacteria are able to infect a wide variety of animal hosts. One model of particular interest is cavitary tuberculosis produced in rabbits by aerosolized virulent tubercle bacilli (Converse *et al.* (1996) Infect Immun **64**(11):4776-4787). In liquefied caseum, the tubercle bacilli grow extracellularly for the first time since the onset of the disease and can reach such large numbers that mutants with antimicrobial resistance may

develop. From a cavity, the bacilli enter the bronchial tree and spread to other parts of the lung and also to other people. Of the commonly used laboratory animals, the rabbit is the only one in which cavitary tuberculosis can be readily produced.

[0065] Vaccines may be formulated according to methods known in the art. Vaccines of the modified bacteria are administered to a host which may be exposed to virulent tuberculosis. In many countries where tuberculosis is endemic, vaccination may be performed at birth, with additional vaccinations as necessary. The compounds of the present invention are administered at a dosage that provides effective immunity while minimizing any side-effects. It is contemplated that the composition will be obtained and used under the guidance of a physician.

[0066] Conventional vaccine strains of BCG may be formulated in a combination vaccine with polypeptides identified in the present invention and produced as previously described, in order to improve the efficacy of the vaccine.

[0067] Various methods for administration may be employed. The formulation may be injected intramuscularly, intravascularly, subcutaneously, *etc.* The dosage will be conventional. The bacteria can be formulated into pharmaceutical compositions by combination with appropriate, pharmaceutically acceptable carriers or diluents, and may be formulated into preparations in semi-solid or liquid forms, such as solutions, injections, *etc.* The following methods and excipients are merely exemplary and are in no way limiting.

[0068] The modified bacteria can be formulated into preparations for injections by dissolving, suspending or emulsifying them in an aqueous or nonaqueous solvent, such as vegetable or other similar oils, synthetic aliphatic acid glycerides, esters of higher aliphatic acids or propylene glycol; and if desired, with conventional additives such as solubilizers, isotonic agents, suspending agents, emulsifying agents, stabilizers and preservatives. Unit dosage forms for injection or intravenous administration may comprise the bacteria of the present invention in a composition as a solution in sterile water, normal saline or another pharmaceutically acceptable carrier.

[0069] The term "unit dosage form," as used herein, refers to physically discrete units suitable as unitary dosages for human and animal subjects, each unit containing a predetermined quantity of vaccine, calculated in an amount sufficient to produce the desired effect in association with a pharmaceutically acceptable diluent, carrier or vehicle. The specifications for the unit dosage forms of the present invention depend on the particular bacteria employed and the effect to be achieved, and the pharmacodynamics associated with each complex in the host.

[0070] The pharmaceutically acceptable excipients, such as vehicles, adjuvants, carriers or diluents, are readily available to the public. Moreover, pharmaceutically acceptable auxiliary substances, such as pH adjusting and buffering agents, tonicity adjusting agents, stabilizers, wetting agents and the like, are readily available to the public.

[0071] The following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to make and use the subject invention, and are not intended to limit the scope of what is regarded as the invention. Efforts have been made to ensure accuracy with respect to the numbers used (*e.g.* amounts, temperature, concentrations, *etc.*) but some experimental errors and deviations should be allowed for. Unless otherwise indicated, parts are parts by weight, molecular weight is average molecular weight, temperature is in degrees centigrade; and pressure is at or near atmospheric.

EXPERIMENTAL

Methods:

[0072] The technical methods used begin with extraction of whole genomic DNA from bacteria grown in culture.

Day 1

[0073] Inoculate culture medium of choice (LJ/7H9) and incubate at 35° C until abundant growth. Dispense 500 µl 1x TE into each tube. (If DNA is in liquid medium, no TE needed.) Transfer loopful (sediment) of cells into microcentrifuge

tube containing 500µl of 1*TE. If taking DNA from liquid medium, let cells collect in bottom of flask. Pipette cells (about 1ml) into tube. Heat 20 min at 80° C to kill cells, centrifuge, resuspend in 500µl of 1*TE. Add 50µl of 10 mg/ml lysozyme, vortex, incubate overnight at 37° C.

Day 2

[0074] Add 70µl of 10% SDS and 10 µl proteinase K, vortex and incubate 20 min. at 65° C. Add 100µl of 5M NaCl. Add 100µl of CTAB/NaCl solution, prewarmed at 65° C. Vortex until liquid content white ("milky"). Incubate 10 min at 65° C. Outside of hood, prepare new microcentrifuge tubes labeled with culture # on top, and culture #, tube #, date on side. Add 550 µl isopropanol to each and cap. Back in the hood, add 750 µl of chloroform/isoamyl alcohol, vortex for 10 sec. Centrifuge at room temp for 5 min. at 12,000 g. Transfer aqueous supernatant in 180µl amounts to new tube using pipetter, being careful to leave behind solids and non-aqueous liquid. Place 30min at -20 C. Spin 15 min at room temp in a microcentrifuge at 12,000g. Discard supernatant; leave about 20µl above pellet. Add 1ml cold 70% ethanol and turn tube a few times upside down. Spin 5 min at room temp in a microcentrifuge. Discard supernatant; leave about 20µl above the pellet. Spin 1 min in a microcentrifuge and discard cautiously the last 20µl supernatant just above the pellet using a pipetter (P-20). Be sure that all traces of ethanol are removed. Allow pellet to dry at room temp for 10 min or speed vac 2-3 min. (Place open tubes in speed vac, close lid, start rotor, turn on vacuum. After 3 min. push red button, turn off vacuum, turn off rotor. Check if pellets are dry by flicking tube to see if pellet comes away from side of tube.) Redissolve the pellet in 20-50µl of ddH2O. Small pellets get 20, regular sized get 30 and very large get 50. DNA can be stored at 4° C for further use.

[0075] *DNA array:* was made by spotting DNA fragments onto glass microscope slides which were pretreated with poly-L-lysine. Spotting onto the array was accomplished by a robotic arrayer. The DNA was cross-linked to the glass by ultraviolet irradiation, and the free poly-L-lysine groups were blocked by treatment with 0.05% succinic anhydride, 50% 1-methyl-2-pyrrolidinone and 50% borate buffer.

[0076] The majority of spots on the array were PCR-derived products, produced by selecting over 9000 primer pairs designed to amplify the predicted open reading frames of the sequences strain H37Rv (ftp.sanger.ac.uk/pub/TB.seq). Some internal standards and negative control spots including plasmid vectors and non-*M.tb.* DNA were also on the array.

[0077] Therefore, with the preparation for an array that contained the whole genome of Mycobacterium tuberculosis, we compared BCG-Connaught to *Mycobacterium tuberculosis*, using the array for competitive hybridization. The protocol follows:

[0078] *DNA labeling protocol.* Add 4 μg DNA in 20μl H_2O, 2 ml dN10N6 and 36 μl H_2O. 2 ml DNA spike for each DNA sample, for total of 60μl. Boil 3 minutes to denature DNA, then snap cool on ice water bath. Add 1 μl dNTP (5mM ACG), 10 μl 10 buffer, 4 μl Klenow, 22 μl H_2O to each tube. Add 3 μl of Cy3 or Cy5 dUTP, for total of 100μl. Incubate 3 hours at 37 C. Add 11μl 3M NaAc, 250 μl 100% EtOH to precipitate, store O/N at -20 C. Centrifuge genomic samples 30 minutes at 13K to pellet precipitate. Discard supernatant, add 70% EtOH, spin 15 minutes, discard sup and speed-vac to dry. This provides DNA for two experiments.

[0079] *DNA hybridization to microarray. protocol.* Resuspend the labeled DNA in 11 μl dH_2O (for 2 arrays). Run out 1 μl DNA on a 1.5% agarose gel to document sample to be hybridized. Of the remaining 10 μl of solution, half will be used for this hyb, and half will be left for later date. Take 5μl of solution Cy3 and add to same amount of Cy5 solution, for total volume 10 μl mixed labeled DNA. Add 1 μl tRNA, 2.75 μl 20x SSC, 0.4 μl SDS, for total volume 14.1μl. Place on slide at array site, cover with 22mm coverslip, put slide glass over and squeeze onto rubber devices, then hybridize 4 hours at 65 C. After 4 hours, remove array slides from devices, leave coverslip on, and dip in slide tray into wash buffer consisting of 1x SSC with 0.05% SDS for about 2 minutes. Cover slip should fall off into bath. After 2 minutes in wash buffer, dip once into a bath with 0.06x SSC, then rinse again in 0.06x SSC in separate bath. Dry slides in centrifuge about 600 rpm. They are now ready for scanning.

[0080] *Fluorescence scanning and data acquisition.* Fluorescence scanning was set for 20 microns/pixel and two readings were taken per pixel. Data for channel 1 was set to collect fluorescence from Cy3 with excitation at 520 nm and emission at 550-600 nm. Channel 2 collected signals excited at 647 nm and emitted at 660-705 nm, appropriate for Cy5. No neutral density filters were applied to the signal from either channel, and the photomultiplier tube gain was set to 5. Fine adjustments were then made to the photomultiplier gain so that signals collected from the two spots containing genomic DNA were equivalent.

[0081] To analyze the signal from each spot on the array, a 14X14 grid of boxes was applied to the data collected from the array such that signals from within each box were integrated and a value was assigned to the corresponding spot. A background value was obtained for each spot by integrating the signals measured 2 pixels outside the perimeter of the corresponding box. The signal and background values for each spot were imported into a spreadsheet program for further analysis. The background values were subtracted from the signals and a factor of 1.025 was applied to each value in channel 2 to normalize the data with respect to the signals from the genomic DNA spots.

[0082] Because the two samples are labeled with different fluorescent dyes, it is possible to determine that a spot of DNA on the array has hybridized to Mycobacterium tuberculosis (green dye) and not to BCG (red dye), thus demonstrating a likely deletion from the BCG genome.

[0083] However, because the array now contains spots representing 4000 spots, one may expect up to 100 spots with hybridization two standard deviations above or below the mean. Consequently, we have devised a screening protocol, where we look for mismatched hybridization in two consecutive genes on the genome. Therefore, we are essentially looking only for deletions of multiple genes at this point.

[0084] To confirm that a gene or group of genes is deleted, we perform Southern hybridization, employing a separate probe from the DNA on the array. Digestions of different mycobacterium DNAs are run on an agarose gel, and transferred to membranes. The membranes can be repeatedly used for probing for different DNA

sequences. For the purposes of this project, we include DNA from the reference strain of *Mycobacterium tuberculosis* (H37Rv), from other laboratory strains, such as H37Ra, the O strain, from clinical isolates, from the reference strain of *Mycobacterium bovis*, and from different strains of *Mycobacterium bovis* BCG.

[0085] Once a deletion is confirmed by Southern hybridization, we then set out to characterize the exact genomic location. This is done by using polymerase chain reaction, with primers designed to be close to the edges of the deletion, see Talbot (1997) J Clin Micro. **35**: 566-9

[0086] Primers have been chosen to amplify across the deleted region. Only in the absence of this region does one obtain an amplicon. PCR products were examined by electrophoresis (1.5% agarose) and ethidium bromide staining.

[0087] Once a short amplicon is obtained, this amplicon is then sequenced. A search of the genome database is performed to determine whether the sequence is exactly identical to one part of the Mycobacterium tuberculosis genome, and that the next part of the amplicon is exactly identical to another part of the Mycobacterium tuberculosis genome. This permits precise identification of the site of deletion.

Below follows an example of the kind of report obtained:

rd6 bridging PCR, blast search of sequence

```
emb|Z79701|MTCY277 Mycobacterium tuberculosis cosmid Y277
Length = 38,908
Plus Strand HSPs:
Score = 643 (177.7 bits), Expect = 1.6e-54, Sum P(2) = 1.6e-54
Identities = 129/131 (98%), Positives = 129/131 (98%), Strand = Plus / Plus
```

```
(SEQ ID NO:130)
Query:    12 ANTAGTAATGTGCGAGCTGAGCGATGTCGCCGCTCCCAAAAATTACCAATGGTTNGGTCA 71
             | ||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
SEQ ID NO:131)
Sbjct:24784 AGTAGTAATGTGCGAGCTGAGCGATGTCGCCGCTCCCAAAAATTACCAATGGTTTGGTCA

Query:    72 TGACGCCTTCCTAACCAGAATTGTGAATTCATACAAGCCGTAGTCGTGCAGAAGCGCAAC
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:24844 TGACGCCTTCCTAACCAGAATTGTGAATTCATACAAGCCGTAGTCGTGCAGAAGCGCAAC

Query:    132 ACTCTTGGAGT 142
              ||||||||||
Sbjct:  24904 ACTCTTGGAGT 24914

Score = 224 (61.9 bits), Expect = 1.6e-54, Sum P(2) = 1.6e-54
Identities = 46/49 (93%), Positives = 46/49 (93%), Strand = Plus / Plus
(SEQ ID NO:132)
Query:    141 GTGGCCTACAACGGNGCTCTCCGNGGCGCGGGCGTACCGGATATCTTAG 189
              | |||||||||||| |||||||| ||||||||||||||||||||||
(SEQ ID NO:133)
Sbjct:  37645 GCGGCCTACAACGGCGCTCTCCGCGGCGCGGGCGTACCGGATATCTTAG 37693
```

[0088] This process is repeated with each suggested deletion, beginning with the three previously described deletions to serve as controls. Sixteen deletions have been identified by these methods, and are listed in Table 1.

[0089] It is to be understood that this invention is not limited to the particular methodology, protocols, formulations and reagents described, as such may, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention which will be limited only by the appended claims.

[0090] It must be noted that as used herein and in the appended claims, the singular forms "a", "and", and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a complex" includes a plurality of such

complexes and reference to "the formulation" includes reference to one or more formulations and equivalents thereof known to those skilled in the art, and so forth.

[0091] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood to one of ordinary skill in the art to which this invention belongs. Although any methods, devices and materials similar or equivalent to those described herein can be used in the practice or testing of the invention, the preferred methods, devices and materials are now described.

[0092] All publications mentioned herein are incorporated herein by reference for the purpose of describing and disclosing, for example, the cell lines, constructs, and methodologies that are described in the publications which might be used in connection with the presently described invention. The publications discussed above and throughout the text are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the inventors are not entitled to antedate such disclosure by virtue of prior invention.